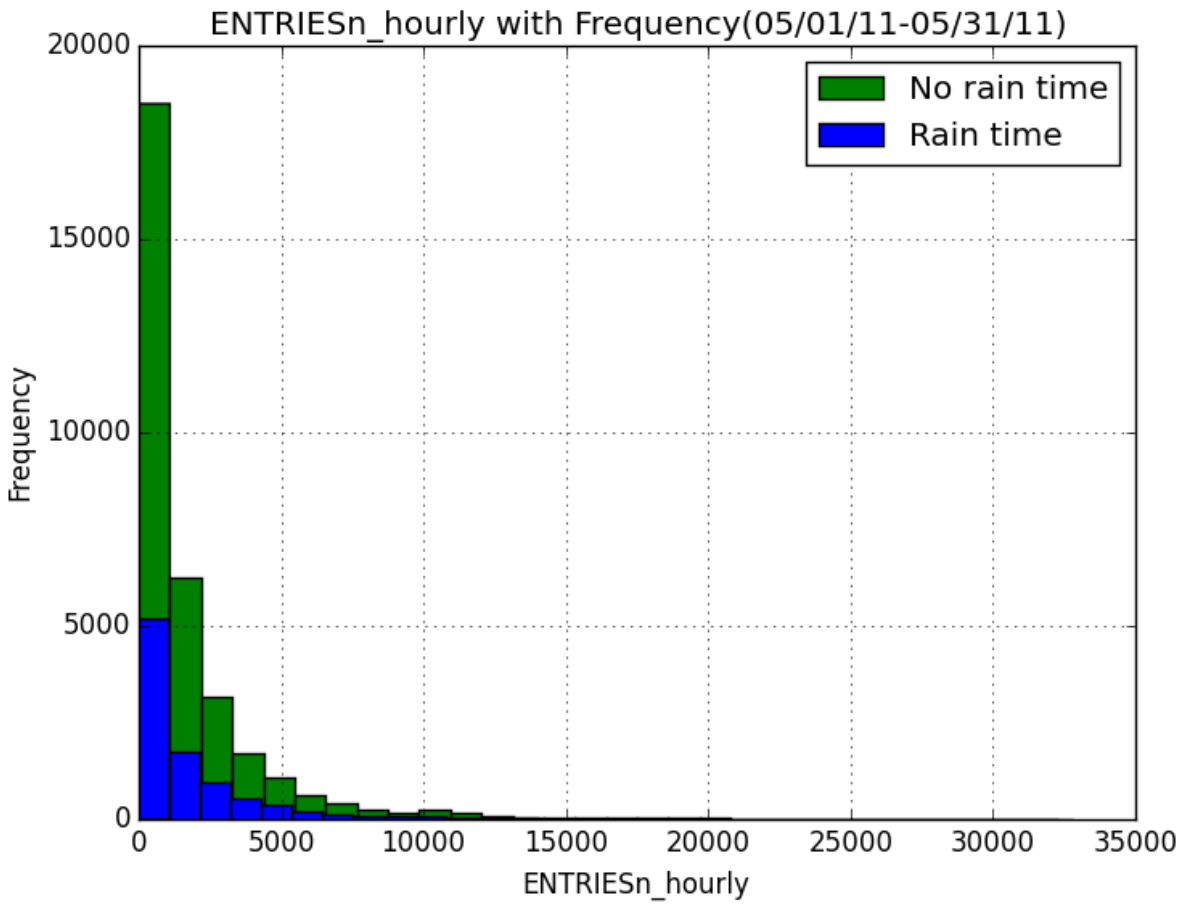
Section 1. Statistical Test

1.1 Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

We want to test if the ENTRIESn_hourly in raining time is statistically significant different from the non raining time or not. If there is a different, how much chances that the different is due to because of the two datasets come from the same population.

Below is the histogram of number of Entries per hour between rain and no rain time with frequency:



The blue bar represents ENTRIESn_hourly in rain time and the green bar represents ENTRIESn_hourly of no rain time. The y-axis is Frequency and the x-axis is the number of entries per hour.

Hence, these two distributions are both positive skew (not normal). We have to use the statistical test that does not assume our data is drawn from any particular underlying probability distribution which is the Mann-Whiney U-Test.

We will use two-tailed test because we do not have any concrete assumption about if the ENTRIESn_hourly of rain time will be higher than no rain time or the ENTRIESn_hourly of no rain time will be higher than the rain time. We also want to test that the different is statistically significant or not or the different is just due to chance.

Our Null hypothesis is that there is no different between the ranks of these two distributions. The Alternative hypothesis is that there is a different between these two distributions whether the ENTRIESn_hourly of rain time is higher than no rain time or the ENTRIESn_hourly of no rain time is higher than in rain time.

The p critical value is 0.05 two tailed test which means to be significantly different the result from the Mann-Whiney U-Test using Scipy has to be less than 0.025 because the result from Scipy is for one tailed test.

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

Because the application of Mann-Whiney U-Test is to test the differences between the two distributions. Which means we are testing if the number of entries per hour in raining time is different from the non raining time or not (in terms of ranking) and if there is a different, how much chances that the different is due to because of the two dataset come from the same population.

Mann-Whiney U-Test is the statistical test that does not assume our data is drawn from any particular underlying probability distribution which is also suitable in our use case because our distribution is positive skew.

By using Mann-Whiney U-Test we want to calculate the chance of having this level of different if two datasets derives from the same population, so the less p value means the less chance of having that level of different if two datasets are from the same population.

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

The p-value = 2.74106957124e-06 = 0.00000274106957124

The mean of Entries per hour of rain = 2028.19603547

The mean of Entries per hour of no rain = 1845.53943866

1.4 What is the significance and interpretation of these results?

The p value from the Mann-Whiney U-Test is 2.74106957124e-06 (0.00000274106957124) which means that there is an extremely statistically significant different between the ranks of entries of rain and no rain distribution at alpha level of 0.05 two tailed test(or even less than alpha level 0.001 two tailed test) which means this level of different between the two datasets is extremely likely not because of the two datasets come from the same population. It might be something else that cause this different to happen.

The mean of Entries per hour of rain = 2028.19603547 and the mean of Entries per hour of no rain = 1845.53943866 which is 2028.19603547 - 1845.53943866 = 182.65659681 Entries per hour different. In conclusion, on average people do ride more subway when it is rain than when it is not rain.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

- Gradient descent (as implemented in exercise 3.5)
- OLS using Statsmodels
- Or something different?

I use Gradient descent.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

I use the improved dataset.

I convert UNIT(station code), day_week(day of the week), conds(Weather condition) and Hour into dummy variable and use it as an input variable.

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

> Your reasons might be based on intuition. For example, response for fog might be: "I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often."
> Your reasons might also be based on data exploration and experimentation, for example: "I used feature X because as soon as I included it in my model, it drastically improved my R^2 value."

I choose to convert UNIT into dummy variable and use it as an input variable because I have an assumption that each subway station will not have the same volume of entries due to differences in its location and differences of population density in that specific area. The UNIT has to be treated as a feature(input variable) otherwise the entries's variability of different station will distract the result of prediction model because there is no variable for explaining this variability.

The result when we perform statistical measurement called R^2 (Coefficient of determination) also correlate with my above assumption. Without UNIT adding into my prediction model, the R^2 value is .1668 which means that my prediction model can only explain 16.68% of variability of ridership, after adding UNIT as a dummy variable, the R^2 value of my prediction model goes up to .5456 which means that my prediction model can better explain variability of ridership up to 54.56% (.3788 increase in R^2 which is quite huge).

I also choose to convert day_week(the day of the week) and hour(Hour of the timestamp) into dummy variable because I have an assumption that in different time of the day and in different day, number of ridership will be different. For example from Monday- Friday in working hour, people might ride more subway than on Saturday and Sunday. In night time of Friday and Saturday people may ride more subway than on Sunday-Thursday.

The result when we perform statistical measurement called R^2 also correlate with the above assumption. Without day_week and hour adding into my prediction model the R^2 value is 0.3813 which means that my prediction model can only explain 38.13% of variability of ridership, after adding UNIT as a dummy variable, the R^2 value of my prediction model goes up to .5456 which means that my prediction model can better explain variability of ridership up to 54.56%(.1643 increase in R^2).

I also choose to convert conds(Weather condition) into dummy variable because I have an assumption that the weather condition might effect number of ridership. For examples in a heavy raining time, number of ridership might increases because people have difficulty driving on the road(the heavy rain makes it difficult for driver to see through the car front panel and the risk of accident increases due to the obstacle of vision and the higher slippery between the tier and the road) (I'm not sure how heavy rain in the US looks like but in Thailand where I live, heavy rain means really heavy.) or in a clear sky time number of ridership might increases due to there is no obstacle to go to subway station and the weather is nice.

The mean of Entries per hour of rain which is 182.65659681 greater than no rain also support my assumption that the weather condition might have some relevance with number of ridership.

The result when we perform statistical measurement called R^2 without conds adding into the prediction model, the R^2 value is .5439. After adding conds as a dummy variable, the R^2 value of my prediction model goes up to .5456 which means that my prediction model explain variability of ridership better only .0017 or 0.17%

The conds cause very less effect on variability of number of ridership compared with other input variable such as hour, day of week and station code. it might be because there is no very serious weather condition which cause people to change their mind about whether they should go to ride subway or not. There is only 288 heavy rain records out of 42649 records and there is also no storm in any records of the dataset which means there might not be any serious weather condition during the time this data has been recorded and that's why the weather doesn't show much correlation with the variability of number of ridership.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?

I do not use any non-dummy variable in my prediction model but to answer this question I will use the coefficient of rain and fog. The dummy variable 'conds' will be taken out of the prediction model only for this question. I use rain and fog column instead. The coefficient result of rain and fog are:

1.86613541e+01 -1.56469062e+01

which means that in my regression model if the rain happen (rain = 1) it will increases the number of ENTRIESn_hourly (compare between rain =0 and rain =1) because the coefficient is positive and if the fog happens (fog = 1) it will decreases the number of ENTRIESn_hourly (compare between fog =0 and fog =1) because the coefficient is negative.

2.5 What is your model's R^2 (coefficients of determination) value?

It's 0.545616156495

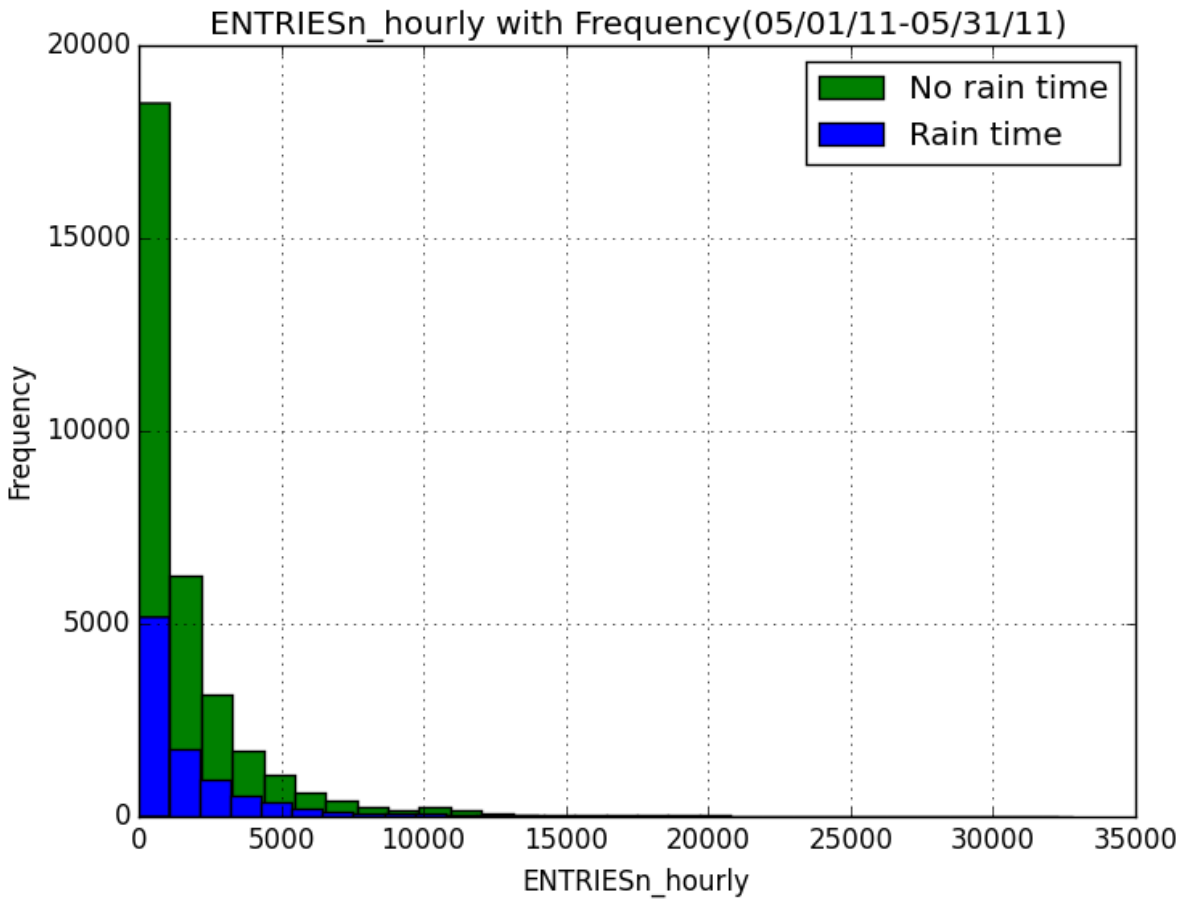
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

It means that 54.56% of the variation in ENTRIESn_hourly can be explained by my regression model and there is 45.44% of variability in number of ridership that my prediction model can't explain.

I think this result of R^2 is appropriate for this dataset because the input variable that we have are quite limited in capability of predicting number of ridership. The varies in the number of ridership is not only due to the weather condition, time, station code, day of week and other variables in the data set. To be more precise in predicting number of ridership we need more input variables such as public holiday, events (concert, big sales event in department store and other public events) in that specific area which might also causes variability in number of ridership.

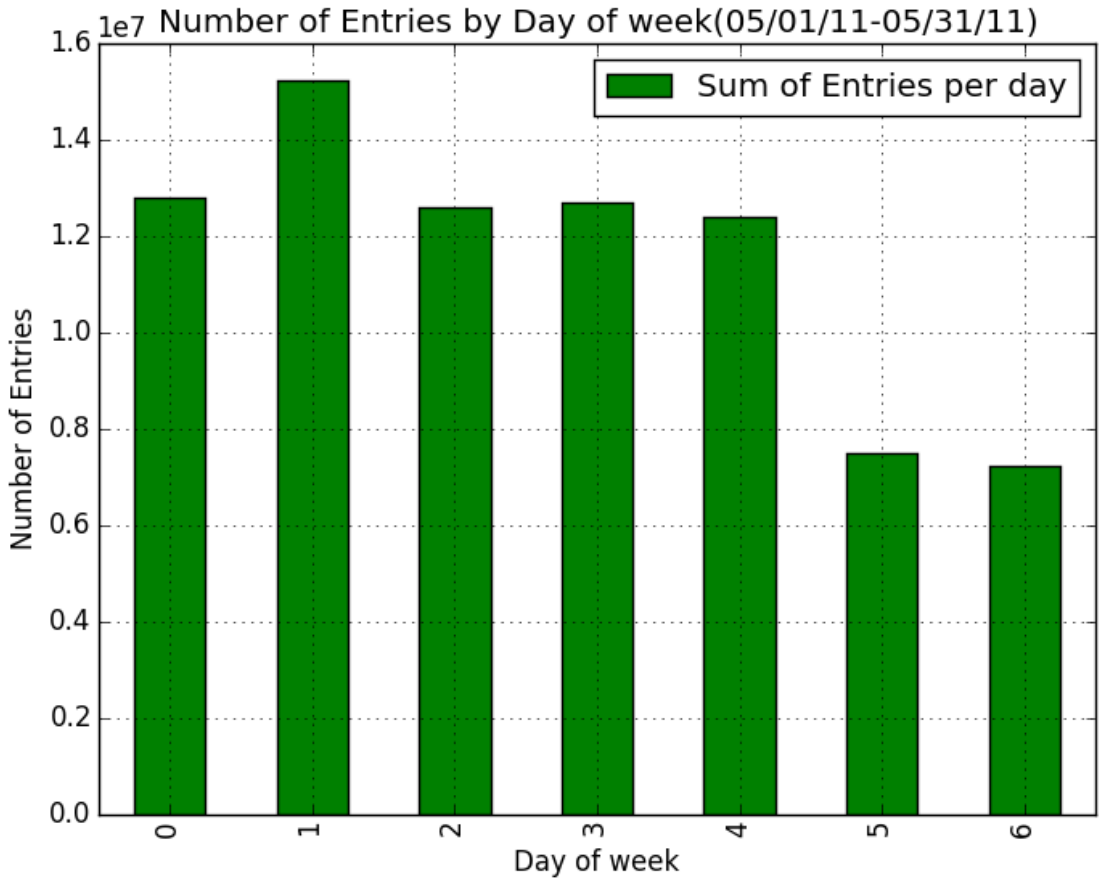
Even through the result of R^2 is appropriate for the dataset but we can make it even better by using other statistical method to improve it such as Ordinary Least Squares (OLS).

Section 3. Visualisation



Above is the Histogram of number of ENTRIEn_hourly with frequency. The Y axis represents Frequency and the X axis represents number of ENTRIESn_hourly. The green bar represents ENTRIESn_hourly of no rain time and the blue bar represents the ENTRIESn_hourly of rain time. The two distribution are positive skew distribution. The mean of ENTRIESn_hourly of No rain time is 1845.53943866 and the mean of ENTRIESn_hourly of rain time is 2028.19603547.

On the next page is the Bar graph of ENTRIESn_hourly grouped by day of week. The number on the X axis is day of week (number 0-6 represent Monday-Sunday). The number on the Y axis is number of Entries (with 1e7). The Green bar graph is the sum of number of Entries between 05/01/11 - 05/31/11 grouped by day of week. The graph shows that from Monday-Friday, number of entries is higher than on Saturday and Sunday.



Section 4. Conclusion

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

I have to conclude that people do ride more NYC subway when it is raining because of the following reasons:

The mean of ridership per hour of rain(2028.19603547) is greater than no rain(1845.53943866) combined with the result of the Mann whitney U test has shown that this different is extremely likely not because of the two datasets come from the same population.

The prediction model also suggest that when there is a rain, the ridership will be more than when there is no rain because the coefficient of rain is 1.86613541e+01 which means that if we compare 2 samples with the exact same features such as same time, same date, same station and etc. but except the first one has

rain=1(rain) and the other one has rain=0(no rain). The one with rain = 1 will have more ridership due to the positive of rain's coefficient.

All the above answer has answered the question "do more people ride the NYC subway when it is raining or when it is not raining?". I would like to add a little more information about whether the rain feature correlate with the variation in number of ridership or not. Below are the additional information which might be out of the scope of the question but I think it is useful:

Having coefficient doesn't imply anything about whether the rain is correlate with variation in ridership or not. If we put anything that irrelevant for predicting number of ridership into the prediction model, we will also get coefficient no matter it relevance with the variation in number of ridership or not. To measure that we have to consider how much R^2 increased after we have added the rain feature into it.

When adding rain feature into the prediction model (remove conds from dataset and use rain and fog feature instead). The R^2 gets very little bit higher (from . 543992 to .544027 which is very little, just .000035 increased) which means that the rain feature helps very little to better describe the variation in number of ridership which means that the rain feature has very little correlation with the variation in number of ridership compared to other features such as day of week, hour, station code.

So in conclusion, people ride more NYC subway when there is a rain than when there is no rain and the different is extremely likely not because of the two datasets come from the same population and the rain feature can be used to help a little bit better describe the variation in number of ridership. The more reason of why the rain can only explains very little bit of variation in number of ridership is above on the answer of question number 2.3 about the conds(because rain is one of the value in conds feature).

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

Dataset,

Analysis, such as the linear regression model or statistical test.

We have to consider about the nature of two datasets before it has been combined. The first dataset is about Subway station the data inside it are such as time, station code, date, number of entries, etc. The second dataset is about weather the data inside it are such as rain, fog, mean temperature, windspeed, etc.

If we want to make prediction model to predict the number of ridership due to the weather data and Subway data, we have to get the entire year of weather data not only 30 days(in the original dataset we only get data from 01/05/2011 - 30/05/2011 and in the improved dataset we only get data from 01/05/2011 - 31/05/2011) because the characteristic of the weather data is base on seasonal ex. summer, winter, raining, etc.

Some variable might not cause high effect until it exceeds some certain value. One variable might have different effect to the variability of number of ridership in different season. For example the mean temperature in winter might cause higher effect on number of ridership than the mean temperature in raining season because when it is below -5 Celsius maybe some people choose to stay at home or leave the city for vacation or it might has too much snow on the road and difficult for car transportation. We never know what it will be or how people will react to the weather condition in that specific situation until we get the data for the entire year and analyse it.

The nature of transportation data is interrelated. The data that we have is one piece of transportation mode of one city (New York city). To make more precise prediction model, the data of all transportation mode has to be integrated with all other related data such as increase in foreigner flight to New York (because New York is one of the global hub, there will be a lot of foreigner there. The changes in number of foreigner will effect the changes in number of ridership), number of take out flight from New York, events in New York, transportation data from other city.

If the goal of this project is to predict the number of ridership. I would say the data provided to this project is quite limited and are not capable enough to make precise prediction. To be more precise in predicting number of ridership we also need more input variables such as public holiday, events (concert, big sales event in department store and other public events) in that specific area and other many variable to be considered.

About shortcoming of the analysis, we didn't calculate confidence interval of parameters (theta). which means that we can not answer the question of "What is

the likelihood we would calculate this parameter value if the parameter had no effect on our output variable?”

We didn’t consider that our data should use linear regression or not. The nature of some datasets might not fit well in Linear regression, sometime we need Non Linear to make more precise prediction.

Last thing, I didn't have time to implement OLS to my prediction model. For Linear regression, using OLS might perform better result than gradient descent because OLS is guarantee to find the optimum solution when perform Linear regression where gradient descent is not.

References

http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_checklist_mannwhitney.htm

http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm

https://www.youtube.com/watch?v=LnfPKGhJypU

https://class.coursera.org/ml-008/

http://stackoverflow.com/questions/28294446/why-i-cant-use-matplotlib-pyplot-in-spyder

http://stackoverflow.com/questions/28212435/how-to-separate-monday-friday-from-saturday-and-sunday-pandas